Validity and Reliability

Kentucky Core Content Test scores are intended for the following uses:

- To provide a basis for state and federal school accountability. Kentucky uses student scores in all tested subjects aggregated to the school level and combined to create the school accountability index for state accountability. It uses the percentage of student scores in reading and mathematics at the level of Proficient or above for federal accountability.

- To support school improvement. School planners use student scores aggregated to the school-by-grade levels to plan and evaluate instructional programs and interventions. Where numbers are sufficiently large, schools may disaggregate scores by gender and/or ethnicity to plan instruction and evaluate programs.

- To inform teachers and parents. For example, schools may include KCCT scores along with other test scores and teacher judgments in placing students and in planning instructional support.

What is a Valid Test Score?

Validity refers to the *meaningfulness* and *appropriateness* of interpretations and actions based on test scores *with respect to their* ***specific use***. In the case of the Kentucky Core Content Tests, the intended uses are as indicated above.

When validity of a test score is discussed in the abstract, there is always an assumption of how the test score will be used. It is important to articulate the use and discuss the validity of the test score in the context of its intended use. For example, one might ask, is a Spanish language test score valid when used to: (1) place a student in a second semester Spanish class or (2) place a student in an English as a Second Language (ESL) class? What might be a valid interpretation in the first instance is not so in the second. A Spanish language test score is not a meaningful or appropriate indicator of how well a student might speak English. It would not be at all appropriate to base an action such as assignment to an ESL class on such a test score. Validity is a use-specific judgment.

Similarly, a test score might be thought more valid for one purpose and less so for another, because validity is a matter of degree, rather than an all-or-none characteristic. A student's ACT score will be helpful to a college in making an admission decision, because it is an indicator, as the publishers say, of the student having learned the academic skills necessary to do well in college. Would the ACT be as appropriate if used to admit a student to military flight training? Perhaps not, since the military has developed its own test, the *Air Force Officer Qualifying Test* with a special section for screening potential pilots.

Like any rational evaluation, a validity judgment must be supported. Evidence must be advanced to support the claim that a test score reflects the construct being tested. Psychometricians, scholars of research methodology and philosophy of social science have contributed a robust body of literature discussing validity and the evidence that can be used to support arguments as to the validity of test scores. The gist is that evidence and argument must be advanced to show that: (a) test scores adequately and representatively reflect the tested construct (e.g., fifth-grade mathematics); (b) test scores are relevant and useful to their intended purpose; c) test score interpretations lead to favorable value implications; and (d) test score uses lead to favorable consequences.

The Kentucky Department of Education has in place an ongoing validation program the purpose of which is to incrementally collect and advance evidence in support of the validity of KCCT test score interpretation and use. Results of the program are summarized and presented to the Kentucky Board of Education and to the Legislative Research Commission as well as published in the biennial KCCT Technical Report and in separate papers. The KCCT Technical Reports and validation papers are available on the KDE website.

What is a Reliable Test Score?

Suppose you are in the habit of weighing yourself on your bathroom scale fairly regularly, almost every morning. Let's say you generally weigh between 115 and 120. One summer you take a vacation to visit your cousin's family, out-of-town. Since you are staying at their house, the next morning you weigh on your cousin's scale. You are horrified to find that you have gained five pounds the very first day. The next day you lose eight pounds. Later that day you lose three pounds, just by taking a shower! It is difficult to say whether you have in fact gained or lost, because the scale readings are inconsistent. Frustrated, you conclude that your cousin's scale is just not as reliable as yours is and that its readings are useless.

Your home scale may not be 100% reliable, but its inconsistencies are tolerable. It may fluctuate by a half pound or so between weighing, but that's about all. An error of a half pound can easily be taken into account. Just add plus or minus a half pound to the scale's reading. However, it is hard to determine your weight using your cousin's scale because it wobbles in both directions over different weighing occasions showing gains and losses of several pounds. Your cousin's scale is just too inconsistent over occasions.

One needs consistency for any measurement to be meaningful. This applies to measurements of student achievement as well as to any other. But how do we determine whether a student test score is reliable? Should we give the student the test over and over on a different occasion? Not a good plan. It is impractical to measure consistency over time on a state test and it is not terribly easy for commercial publishers to do so.

It is practical, however, to measure consistency over test items. A large number of test items are featured on each form of the KCCT. This allows us to measure the KCCT's internal consistency (from task to task). For example, the 2005 fifth-grade mathematics test has 24 multiple-choice and six open-response items. The reliability statistic computed over items is referred to as Cronbach's alpha or coefficient alpha. Reliability statistics are based on that notion that perfect reliability would be equal to 1.00. When coefficient alpha is high, about .80 or above, reliability is high. When it is low, about .50 or below, we don't have a great deal of confidence in the internal consistency of the test. Higher test reliability is associated with larger numbers of test items. The table below presents median coefficient alpha by Core Content Test and Year.

Table 14-2
KCCT Reliability 2000 – 2004
Median and Range Coefficient Alpha Across Forms* by Grade and Subject

Core Content Test by Grade		2001		2002		2003		2004		2005	
		Median[1]	Range	Median[1]	Range	Median[1]	Range	Median[1]	Range	Median[1]	Range
4/5	Reading	.88	.87-.88	.88	.86-.88	.86	.85-.87	.87	.84-.87	.86	.85-.89
	Mathematics	.87	.86-.88	.87	.86-.88	.87	.86-.87	.86	.83-.88	.87	.84-.88
	Science	.84	.80-.85	.83	.81-.84	.83	.82-.85	.84	.81-.85	.82	.81-.83
	Social Studies	.84	.84-.85	.85	.83-.86	.84	.83-.85	.83	.82-.84	.83	.81-.84
	Arts & Hum	.66	.63-.67	.66	.63-.71	.66	.62-68	.65	.61-.73	.64	.60-.67
	PL/VS	.63	.53-.67	.69	.67-.73	.61	.50-.64	.59	.49-.67	.58	.53-.62
7/8	Reading	.87	.87-.88	.87	.87-.88	.86	.85-.87	.86	.85-.87	.86	.85-.86
	Math	.89	.88-.90	.89	.88-.90	.89	.88-.89	.89	.88-.90	.89	.88-.90
	Science	.84	.83-.86	.86	.84-.86	.85	.84-.86	.85	.84-.86	.85	.83-.87
	Social Studies	.89	.87-.89	.88	.87-.89	.88	.87-.89	.88	.86-.88	.87	.85-.88
	Arts & Hum	.70	.66-.73	.69	.67-.73	.67	.59-.73	.66	.61-.73	.68	.63-.72
	PL/VS	.70	.66-.74	.71	.67-.74	.68	.63-.73	.67	.62-.71	.66	.63-.72
10/12	Reading	.88	.87-.89	.88	.88-.89	.87	.87-.88	.89	.88-.91	.89	.87-.89
	Mathematics	.88	.85-.89	.89	.87-.89	.89	.88-.89	.89	.88-.90	.88	.88-.90
	Science	.84	.82-.85	.85	.81-.86	.84	.82-.85	.84	.82-.84	.84	.84-.85
	Social Studies	.88	.87-.88	.89	.88-.89	.88	.87-.89	.88	.87-.89	.88	.87-.89
	Arts & Hum	.67	.61-.72	.69	.65-.72	.66	.62-.68	.67	.57-.71	.66	.60-.72
	PL/VS	.64	.60-.68	.65	.62-.68	.64	.56-.67	.63	.57-.71	.62	.51-.65

[1]Median coefficient alpha is based upon operational matrix OR and MC items across test forms.
Six test forms are used in reading, mathematics, science and social studies; twelve forms in Arts and Humanities and Practical Living/Vocational Studies.

The 2005 KCCT Technical Report Appendices present coefficient alpha for each Kentucky Core Content Test in each Core Content from 2001 to 2005. Note that coefficient alpha must be computed separately for each test form, since each form of the KCCT has different items. The range of coefficient alpha statistics in a Core Content, as well as the median coefficient alpha (middle), is presented in Table 14 – 2 above. Median coefficient alpha for fifth-grade mathematics in 2005 is .87 and the range is .84 to 88. Notice that coefficient alpha statistics for all subjects (except Arts and Humanities and Practical Living/Vocational Studies) range between values of .80 and .90. The two tests showing lower reliability coefficients (.53 and .72) have fewer

numbers of items. It is recommended that student-level interpretations take this into account.

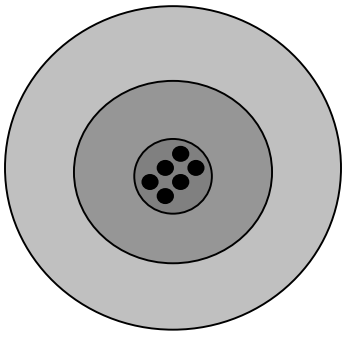
Which is more important reliability or validity? Both are important.

Can a test score be valid if its reliability is low? Lower levels of reliability weaken the case for the valid test score test score interpretation. Moderate levels of reliability, such as those shown for the Arts and Humanities and Practical Living Vocational Studies Tests, are cause for caution, when test scores are interpreted at the student level.

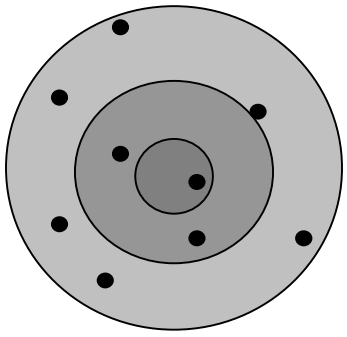
Can a test score be reliable, but invalid? Yes. A test can be reliable, but invalid for a specific use, as in the example mentioned above, when it was suggested that a Spanish language test be used to place student in an ESL program.

Think of it this way. A valid test score is on target. A reliable test score is consistent. A valid and reliable test score is on target and consistent. These relationships are illustrated in the graphic below.

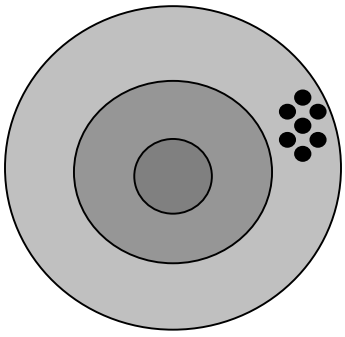
Note: Numerous studies on the validity and reliability of CATS have been performed and a list of those studies is attached. Copies of these studies are available upon request.



Reliable and Valid



Unreliable + Invalid



Reliable + Invalid